A BENCHMARK FOR SPEAKER-DEPENDENT RECOGNITION USING THE TEXAS INSTRUMENTS 20 WORD AND ALPHA-SET SPEECH DATABASE

David S. Pallett

Institute for Computer Sciences and Technology
National Bureau of Standards
Gaithersburg, MD 20899

ABSTRACT

This paper presents the results of performance assessment tests conducted on one commercially available speaker-dependent template-matching speech recognizer, using a widely available speech database. Test vocabularies include the Texas Instruments 20 word test vocabulary and the 26 letters of the spoken English alphabet (the alpha-set). For the 20 word set, overall recognition accuracy was 99.24%, and for the alpha-set it was 84.88%. Comparisons are made with the performance of research systems which use both template matching and feature-based technologies, as well as with the results of tests on commercially available recognizers of 5-7 years ago. The intended purpose of these measurements is to provide a benchmark for comparing the results of tests of more sophisticated systems.

INTRODUCTION

As the performance of speech recognition technology improves, more challenging test material is required in order to demonstrate the capabilities of improved systems. For speaker-dependent isolated word recognition, widespread dissemination and use of the 20 word Texas Instruments (TI) speech database (first used in Doddington and Schalk's study of the state-of-the-art in 1981 [1]), has provided a valuable research resource and measures of performance that serve as benchmarks for this 20 word vocabulary. However, as performance of the technology has improved, the value of this database has declined because it may no longer provide substantial challenge to the current state-of-the art. More challenging speech test vocabularies and databases are required in order to demonstrate improved capabilities.

In large vocabulary natural language systems, the spoken letters of the English alphabet, or the "alpha-set", may be widely used to introduce the spelling of new words in the lexicon. This application has been termed "spellmode". In such an application, the use of syntax to restrict the vocabulary is obviously inappropriate, and the required use of special-purpose alphabets such as the International Civil Aviation Phonetic Alphabet is probably undesirable. The use of the alpha-set is natural in such an application.

At the time that the TI 20 word database was collected, the same talkers also provided tokens for the alpha-set [2], and this speech database is now in the public domain. The availability of this test material provides a means for comparative tests on both the 20 word database and the alpha-set for the same set of talkers, and increases the value of the original 20 word database by providing more challenging material from the same group of test talkers that was obtained under identical environmental conditions.

This paper presents preliminary results on tests of performance on the TI 20 word vocabulary and the alpha-set for a representative commercially available speaker-dependent recognizer costing approximately $1000. These data are intended to provide benchmarks of performance for comparison of the performance of more sophisticated recognition algorithms, using speech database material that is widely available. More detailed analysis of this data is being conducted and at least two other commercially available recognizers are to be studied.

TEST PROCEDURES

The tests reported upon in this paper were conducted using procedures outlined in a recent paper [3]. They reflect suggestions on experimental design, data analysis and documentation from the IEEE Speech I/O Technology Performance Evaluation Working Group. Material included in this section follows the format suggested in this reference.

Experimental Design

These tests were intended for benchmark purposes. The TI 20 word vocabulary and the alphabet were used in separate tests, with no use of syntax to control the active recognition vocabulary. The use of these vocabularies may be representative of an application such as "spellmode", but no explicit effort is taken to model an application.

Test Talker Population

Eight males and eight females comprise the test talker population, with no effort taken to control dialect.

Test Vocabulary

The 20 word vocabulary consists of the words "yes, no, erase, rubout, repeat, go, enter, help, stop, start" and the digits "zero" through "nine". The alpha-set consists of the letters "a" through "z". All words were spoken as discrete utterances. It is interesting to note that the available tokens in the database could be recombined to yield an "alphadigit" set as used in other studies [4,5], but this study sought to direct attention to a comparison of performance for the 20 word and alpha-sets.

Training

The database includes 10 tokens of each of the 46 words for each talker. These tokens are intended for use in training or enrollment. This material was used for enrollment in accord with the manufacturer's recommendations. Typically, the first token was used for 'enrollment', and three additional tokens were used to 'update' the resulting reference patterns or templates. Training was implemented automatically, and no attempt was taken to optimize the reference template set.

Environment

Test material was obtained in a quiet sound-isolation booth with a cardiod dynamic microphone placed approximately 2 inches from the talker's mouth. The speech signal-to-noise ratio is believed to exceed 40 dB, but (to date) has not been measured.

Recorded Test Material

The speech signal was initially digitized with a 12-bit A/D converter at a 12.5 kHz sampling rate. The digital data were made available to the National Bureau of Standards by Texas Instruments for use in the public domain. An analog signal was reconstructed using a D/A converter, using a 6.3 kHz antialiasing filter. This audio signal was then recorded using commercially available PCM/VCR technology with a digital mastering processor and a video cassette recorder.

One audio channel on the PCM/VCR recorded material provides a recorded modem signal with ASCII character string data that precedes each utterance recorded on the other audio channel. The use of this format and 'header' data facilites automatic enrollment and scoring [6].

Playback of the recorded material provides two line-level audio signals, one for the modem and one with the test material. The line-level audio signal with the test material was used as input to a mixer, with the microphone level output of the mixer used as input to the recognizer. Headphones driven by the mixer were used to monitor the signal as desired.

Calibration tones provided on the PCM/VCR recorded material were used to establish system gains, and tests were conducted using the recognizer manufacturer's routines to establish appropriate recognizer gains. Once gains were established, they were fixed, and no effort was taken to optimize gains for improved performance.

Statistical Considerations

There are a total of 5120 test tokens for the 20 word vocabulary (16 test tokens for each of the 20 words for each of the 16 talkers). There are a total of 6655 valid test tokens for the alpha-set. One test token of one letter ("s") for one talker (f5) has been found to contain only breath noise. For each of the 16 talkers, there are 10 training tokens available for each of the 46 words in the two vocabularies of the database, for a total of 7360 training tokens. The total number of tokens in the database is thus 19135 tokens.

Since the total number of errors per talker is small, the precision associated with these data is unknown.

Several repetitions of tests for individual talkers were conducted in order to assess the variability between repeated tests. These tests included repeated enrollment and repeated use of the of the test material on a given template set. In general, there has been very good repeatability, typically varying in one count of the total number of substitutions. Although the number of errors per talker is larger for the alpha-set, the variability in the count of the total number of

substitutions is typically three or four.

BENCHMARK DATA

20 Word Vocabulary

Overall Scores for 5120 tokens
for 8 males and 8 females

Correct Recognition Percent:	99.24%	(5081)
Substitution Percent:	0.61%	(31)
Deletion Percent:	0.06%	(3)
(No Insertions)		
Rejection Percent:	0.10%	(5)
Ratio of total errors to rejections:	6.8	

Figure One indicates the distribution of responses for the 20 word vocabulary. In this matrix representation, the input words are listed along the rows, and the recognizer's responses are shown in the appropriate column.

Alpha-set

Overall scores for 6655 valid tokens
for 8 females and 8 males

Correct Recognition Percent:	84.88%	(5649)
Substitution Error Rate:	14.92%	(993)
Deletion Percent:	0.03%	(2)
(No Insertions)		
Rejection Percent:	0.17%	(11)
Ratio of total errors to rejections:	90.4	

Figure Two indicates the confusion matrix for the alpha-set.

The intended test procedure was to disable the reject capability of the recognizers under test to facilitate comparisons. For this recognizer, it was not possible to do so. Following the manufacturer's recommendation, the acceptance threshold was set to its maximum value, and no restrictions were imposed on the 'closeness' of best and next-best scores. This results in a very low, but non-zero rejection percent. Performance on the alpha-set might be improved by the imposition of appropriate reject criteria.

For the 20 word set, there are 5 rejections for the male talkers and no deletions, and no rejections and 3 deletions for the female talkers. For the males, recognition accuracy and substitution error percents are, respectively, 98.94% and 0.86%, with corresponding data for the females 99.53% and 0.35%.

For the alpha-set, there are 10 rejections for the male talkers and no deletions, and 1 rejection and 2 deletions for the female talkers. For the males, recognition accuracy and substitution error percents are, respectively, 83.4% and 16.3%, with corresponding data for the females 84.9% and 14.9%.

DISCUSSION

In comparing error rates for the two test vocabularies, the substitution error rate is approximately 20 times larger for the alpha-set, reflecting the greater difficulty of recognizing a vocabulary consisting exclusively of monosyllables (with the sole exception of "W"), containing several highly confusable subsets, and with a branching factor that is 30% larger.

The small number of substitution errors observed for the 20 word vocabulary (31) is believed to fairly represent the state-of-the-art of currently available low-cost recognizers. Further data are to be obtained on other recognizers, including the use of other approaches including stochastic modelling. By comparison with the results of Doddington and Schalk's 1981 benchmark data, the error rate is half that of the second-best recognizer in their tests (a template matching unit then having a nominal price of $65,000).

As previously noted for the TI 20 word data base [1], there is considerable variation between individual talkers' scores. For the 20 word set, individual scores range from 97.5% to 100%, while for the alpha-set the range is from 74.3% to 91.8%. In general, "sheep" and "goats" retain their relative rank-order places when comparing results for the two vocabularies. These variations underscore the need for adequate population sampling and large enough test data bases for statistical validity.

In studying the confusion matrix that results from separate consideration of members of the E-set as input (Figure Three), it is evident that the bulk of the substitution errors occur for the E-set (the letters "B,C,D,E,G,P,T,V" and "Z"). There are a total of 2304 test tokens in this subset. There are a total of 1546 correct responses and 754 substitution errors, for a subset recognition accuracy of 67.1% and a substitution percent of 32.7%. The 754 substitution errors for the 9 word E-set comprise approximately 75% of all substitution errors for the 26 word alpha-set. Approximately 98% of all substitution errors for E-set input tokens

fall within the E-set.

Within the E-set, overall recognition accuracies for individual letters range from 92.9% for "E" to 53.1% for "D", with significant variations occurring for different talkers.

Previous measures of the ability of template matching systems to perform fine phonetic distinctions as cited in Cole et al. [7] indicate recognition accuracy for the E-set at about 60%. The present measurements on a commercial product suggest slightly better performance, with considerably better performance for "E" (92.9%) and "G" (90.6%) than for other members of the subset such as "B" (58.6%) and "D" (53.1%).

In Cole's work comparing template matching and feature based recognition, an alpha-set data base of 2080 tokens was used (4 tokens of each letter produced by 10 female and 10 male talkers). The system under study was operated in a speaker-independent mode, with a procedure used to ensure that the test talker's data were consistently deleted from the training material. Without tuning (adaptation to individual talker's speech), an overall error rate for the alpha-set of 10.5% was obtained, in contrast with the error rate of 14.92% found for the speaker-dependent recognizer in this study.

For the E-set, Cole cites an error rate of 14% in contrast with the 32.7% in this study. Using tuning (on the limited number of tokens available for each letter for each speaker in his data base) and improved algorithms, Cole indicates that an error rate of 6% was obtained, approximately one-fifth that of this commercially available template-matching recognizer. This comparison suggests the strength of the speaker-independent feature-based recognition technology when compared with current technology. Further comparisons with the performance of speaker-dependent (or adaptive) systems using stochastic models should be informative.

SUMMARY

This paper reports on preliminary tests conducted using a widely available speech data base in the public domain and a commercially available recognizer using template matching technology. For the 20 word vocabulary used in these tests, a recognition accuracy of 99.24% was measured, while for the spoken English alphabet, recognition accuracy was 84.88%. The 9 members of the E-set are responsible for 75.9% of all substitution errors for the 26 letters of the spoken alphabet.

These tests suggest that the performance of current low-cost commercial products using template matching technology is slightly superior to results reported for research systems of 5 to 7 years ago and to that of commercially available systems coating as much as $65,000 of that era. The tests also suggest that performance is inferior to more sophisticated systems using stochastic modelling and/or acoustic-phonetic feature based recognition.

REFERENCES

[1] G.R.Doddington and T.B.Schalk, "Speech Recognition: turning theory to practice", IEEE Spectrum, September 1981, pp.26-31.

[2] T.B.Schalk, "The Design and Use of Speech Recognition Data Bases", Proceedings of the Workshop on Standardization for Speech I/O Technology, D.S.Pallett, (ed.), National Bureau of Standards, March 1982.

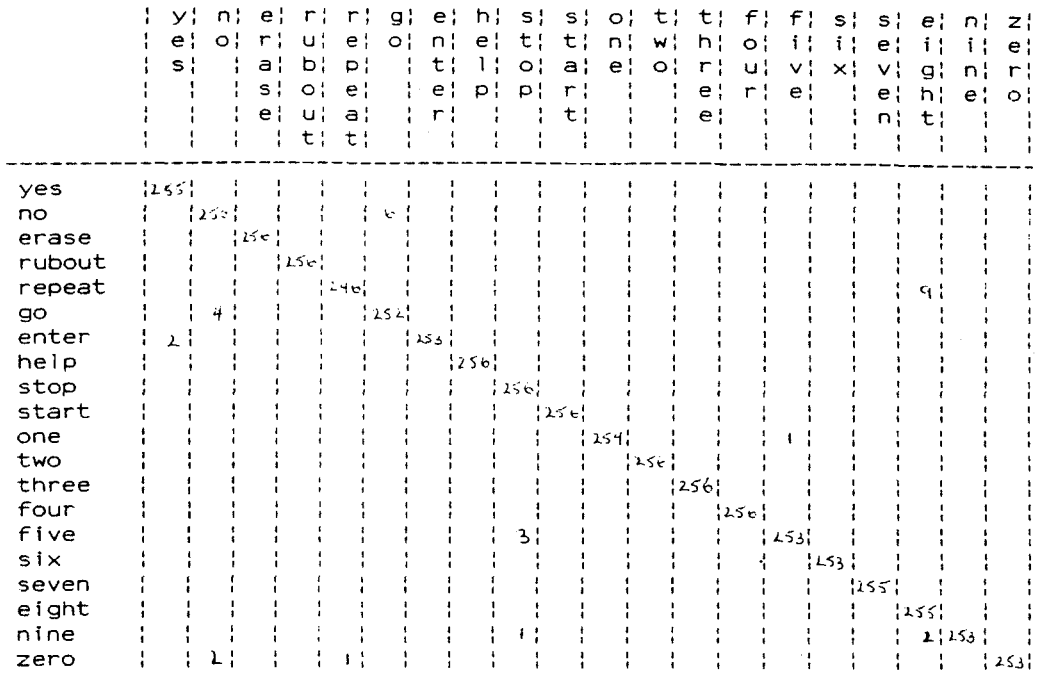
[3] D.S.Pallett, "Performance Assessment of Automatic Speech Recognizers", Journal of Research of the National Bureau of Standards, Vol. 90, No. 5, September-October 1985, pp.371-387.

[4] N.R.Dixon and H.F.Silverman, "What are the significant variables in dynamic programming for discrete utterance recognition?", Proceedings of ICASSP'81 (Atlanta), pp.728-731.

[5] L.F.Lamel and V.W.Zue, "Performance Improvement in a Dynamic-Programming-Based Isolated Word Recognition System for the Alpha-Digit Task", Proceedings of ICASSP'82 (Paris), pp.558-561.

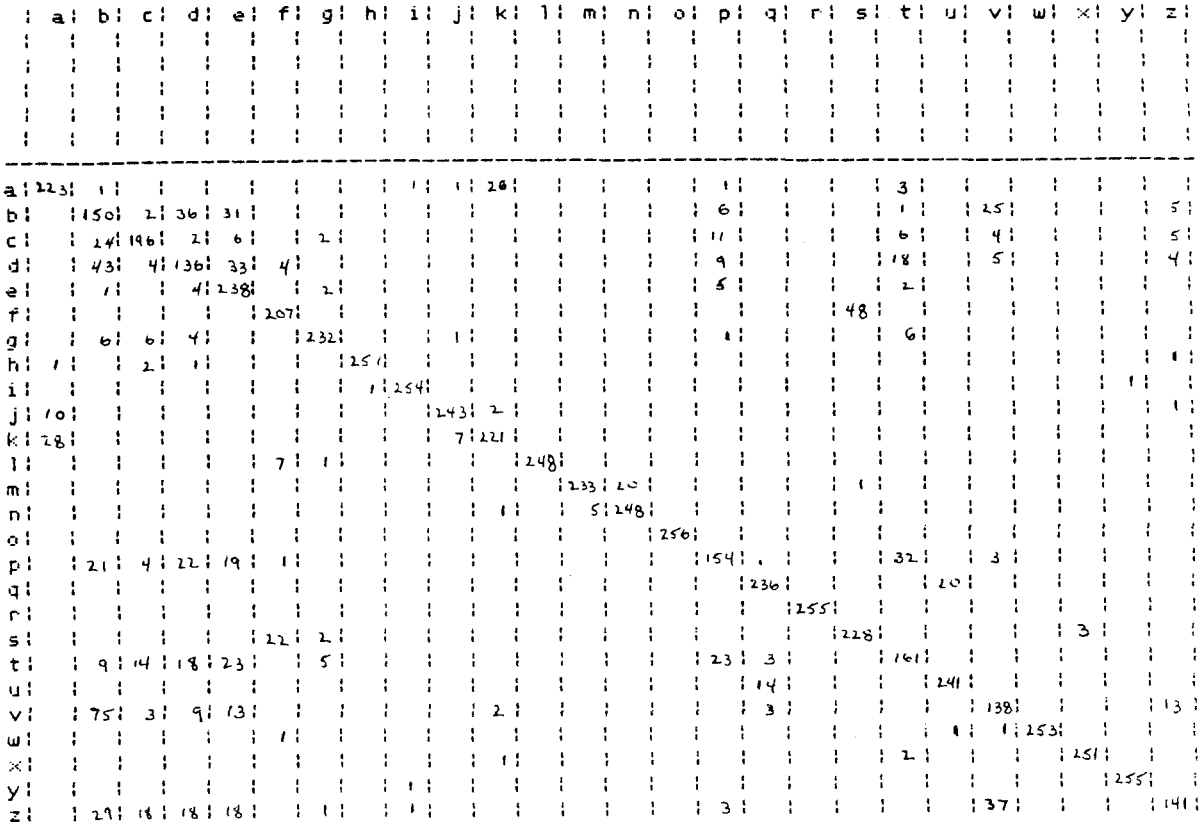
[6] D.S.Pallett, "A PCM/VCR Speech Database Exchange Format", to appear in Proceedings of ICASSP'86 (Tokyo).

[7] R.A.Cole, R.M.Stern and M.J.Lasry, "Performing Fine Phonetic Distinctions: Templates vs. Features", in Conference Record of "Toward Robustness in Speech Recognition", W.A.Lea, (ed.), Santa Barbara, CA November, 1983.



	yes	no	erase	rubout	repeat	go	enter	help	stop	start	one	two	three	four	five	six	seven	eight	nine	zero
yes	255						2													
no		255				6														
erase			256																	
rubout				256																
repeat					246													9		
go		4				252														
enter	2						253													
help								256												
stop									256											
start										256										
one											254				1					
two												256								
three													256							
four														256						
five									3						253					
six																253				
seven																	255			
eight																		255		
nine									1									2	253	
zero		2			1															253

Figure One: Confusion matrix representing responses for the 20 word vocabulary.



	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	223	1							1	1	26					1				3						
b		150	2	36	31											6				1		25				5
c		24	196	2	6		2									11				6		4				5
d		43	4	136	33	4										9				18		5				4
e		1		4	238		2									5				2						
f						207													48							
g		6	6	4			232			1						1				6						
h	1		2	1				254																		1
i								1	254																1	
j	10									243	2															1
k	28									7	221															
l						7	1					248														
m													233	20					1							
n											1		5	248												
o															256											
p		21	4	22	19	1										154				32		3				
q																	236				20					
r																		255								
s						22	2												228					3		
t		9	14	18	23		5									23	3			161						
u																	14				241					
v		75	3	9	13						2						3					138				13
w						1															1	1	253			
x											1									2				251		
y									1																255	
z		29	18	18	18		1		1							3						37				141

Figure Two: Confusion matrix representing responses for the alpha-set.

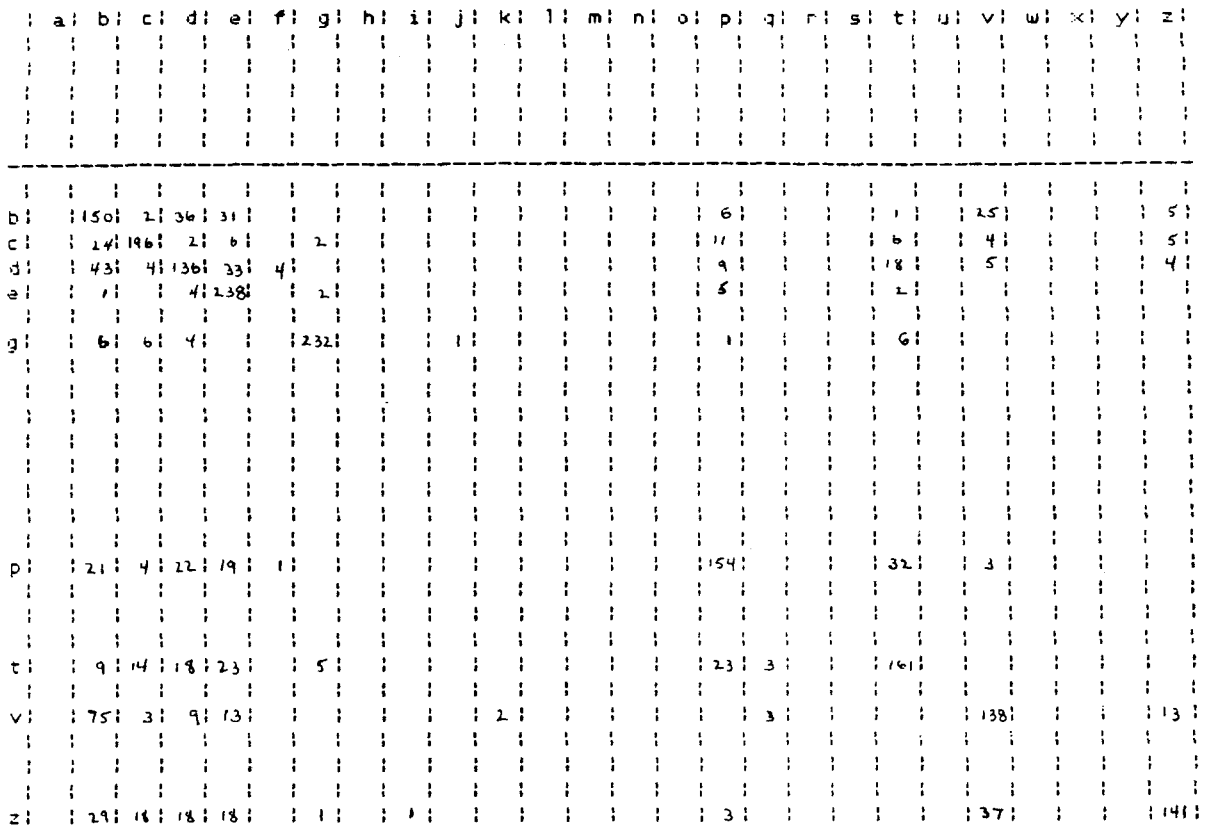


Figure Three: Confusion matrix representing responses for the E-set.